ANALYZING ITEM PARAMETER ESTIMATES IN MATHEMATICS MULTIPLE-CHOICE ITEMS OF THE NATIONAL EXAMINATION COUNCIL USING 2-PARAMETER MODEL

[1]FALEYE, F. O., [2]BABATIMEHIN, T., [3]OGUNGBAIGBE., T. S. & [4]AJEIGBE, T. O.

[1]School of Education, Adeyemi Federal University of Education, Ondo

[2,3&4]Department of Educational Foundations & Counselling, Obafemi Awolowo University, Ile-Ife.

Omowumi_faleye@yahoo.com, tbabatimehin@oauife.edu.ng, tsogungbaigbe@oauife.edu.ng, taiaje@oauife.edu.ng

Abstract

The study analyzed the item parameter estimates of the Mathematics multiple-choice items. It also determined the reliability of the items, as well as the item characteristic curve of the items. These were with a view to providing empirical statistical evidence on the nature of the items used by National Examinations Council. An ex-post-facto research design was adopted for the study. Senior secondary three students who sat for SSCE Mathematics examination in Osun State during the 2012 constituted population for the study. A sample of 2500 was selected through stratified sampling technique, using school location and sex as stratum. Data collected was based on responses to 60 Mathematics multiple-choice as contained in the scanned Optical Mark Record (OMR) sheets. Analysis of the data was carried out X-Calibre 4.2. The results of the item difficulty revealed that 16(26.7%), 22(36.7%) and 22(36.7%) of the items were easy, moderately, and difficult respectively; in terms of discrimination indices, 23(38.3%), 13(21.7%), and 24(40.0%) had poor, marginal, and moderate discriminations respectively. The result also revealed reliability coefficient of 0.85 for the calibrated 60 Mathematics multiple-choice items, which implied appropriate internal consistency. The results finally revealed that 32 (55%) of the 60 items complied with the assumption of IRT with reference to Item Characteristics Curve under the 2-parameter model. The study therefore concluded that the Mathematics multiple-choice items used by NECO were moderate in terms of quality

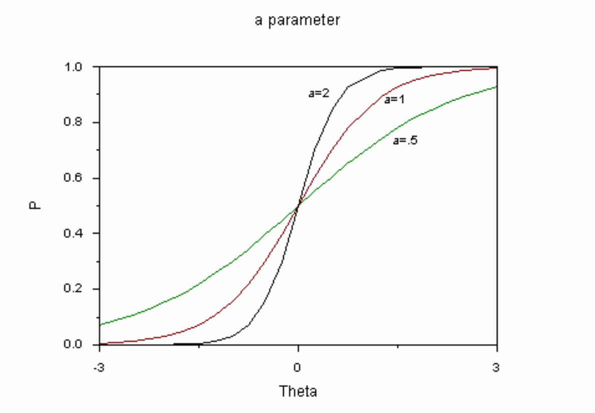
Keywords: Item, Parameter Estimates, Multiple-Choice, National Examination Council and 2-Parameter Model

Introduction

In the world of Education, the evidence of teaching and learning is the resultant learning outcome. Learning outcomes are used in making educational decisions such promotion, placement, selection, remediation etc about learners. For such decisions to have merit, the process should be adjudged valid and reliable. In testing, items use to assess students' ability should be free of measurement error. To establish or ensure that test items are error free, a modern approach know as Item Response Theory (IRT) is proposed as against the Classical Test Theory (CTT). The IRT is regarded as a better selection statistical approach for test constructors with greater flexibility and stable parameter estimates (Ojerinde, Popoola, Ojo, & Onyeneho, 2012). The IRT model comprised models estimate the probability of a given response based on additional item characteristics such as discrimination and guessing (Bond & Fox, 2001). The Two parameter model (2PL) additionally estimates item discrimination (a). The *a* parameter is found by taking the slope of the line tangent to the ICC at b. The *a* parameter is the steepness of the curve at its steepest point. The a parameter is called the discrimination parameter. Its closest relative in classical test theory is the item total correlation. The item response function (IRF) is written as:

$$Pi(\Theta) = \frac{1}{1+\exp[-1.7\alpha(\Theta-b)]}$$

Where: Pi(Θ)the probability that an examinee with ability level Θ answers item I correctly; b= the item difficulty parameter, a= the item discrimination parameter, 1.7= scaling factor (D)



The steeper the curve, the more discriminating the item is, and the greater its item total correlation. As a limit, a step function can be set where below some level, the probability of getting the item right is zero, and just above that, the probability jumps to 1.0. As the *a*

parameter decreases, the curve gets flatter until there is virtually no change in probability across the ability continuum. Items with very low *a* values are not good for distinguishing among people, just like items with very low item total correlations. The 2-parameter model allows both *a* and *b* parameters to vary to describe the items. This model is used to represent attitude scales and some ability tests where there is no guessing. From purely statistical considerations, test construction using CTT might often consist of selecting those items with the best discrimination (item-total correlation) and which span a range of item difficulties (Mead & Meade, 2010). The question is “is the difficulty of the subject resides in its self”? As such, a clear understanding of Mathematics will prepare students for future challenge in the subject, especially at the tertiary level of education. The subject is one of the important subjects students are expected to pass at the credit level, to avail them the opportunity to study science based courses (Medicine, Pharmacy, Botany, Nursing etc) at the University level. Students, bearing this in mind would have prepared to succeed in the subject. The extent to which student are ready for the next phase of learning needed to be examined (Klein and Hamilton 1999). However results released by public examinations bodies in Mathematics over the years have not been encouraging. Psychometrians are more concerned with problems measurement error that are like to affect the item quality. This may be connected to the fact that the test instruments should be yield a valid and reliable measure of students' ability. z importance associated with measurement Item characteristic curve had been found useful in selecting quality items Thorpe and Favis (2012). It was reported that ICCs had empirical strength with 95 percent of error free.

In a school system, assessment is an integral aspect of teaching and learning without which it may be difficult to ascertain the extent to which learning as taken place. Its usefulness is not only limited to measuring learning outcome alone, it is also a means of judging the adequacy of teaching techniques. Assessment as gone beyond classroom teachers, several examination bodies such West Africa Examination Council (WAEC), National Examinations Council (NECO), National Technical Examination Board (NATEB) are by law in charge of private and public examinations in Nigeria. This implied that the general public are watching with kin interest in the feedback from these various examination. This may be traceable to the fact that students' results emanate from them, thereafter are used as means of getting admission into the tertiary institutions in the country. In view of this, one will expect that test items use in assessing students' cognitive ability is adequate in terms of being valid and reliable.

Mathematics is one of the major subjects that cross across every levels of Education, from primary to tertiary level. Its relevance, especially in advancement of technology cannot be overstressed. The operations in Mathematics are performed by students during their daily activities, making it an important aspect of human life (Gocken, 2014). Students from various backgrounds are expected to perform the same tasks in Mathematics having been exposed to series of learning instruction. Often times there have been complains about low performance of students in Mathematics. Low in performance may be traced to various factors such; students' disposition, teachers' competence, availability of teaching and leaning materials, item structure and so on. This paper tailored towards item structure with respect to its quality. However, there is the need to apply a modern test theory (Item Response Theory) to ascertain the quality of the Mathematics multiple-choice items, being one of the major relevant subjects students needed to pass at credit level to study related science-based courses in any tertiary institutions in Nigeria, hence this study aims to:

a) analyse the item parameter estimates of the Mathematics multiple-choice items using item response approach;
b) determine the reliability of the Mathematics multiple-choice items; and
c) examine the extent to which Mathematics multiple-choice items comply with the assumption of IRT with reference to item characteristic curve under 2-parameter model.

Research Questions

1. What are the item parameter estimates of the Mathematics multiple-choice items using 2-parameter?
2. What is reliability coefficient of the Mathematics multiple-choice item?
3. To what extent do Mathematics multiple-choice items comply with the assumption of IRT with reference to item characteristic curve under 2-parameter model?

Methodology

Ex-post-facto research design was adopted for the study. This design fits into the study because the data was pooled from data base of candidates' responses in the 2012 NECO Mathematics test items, which already being administered and scored. The population consisted of 13,355 candidates, out of which 6748 (50.5%) and 6606 (45%) boys and girls respectively. An intact class of 2500 candidates was selected as sample from 36 randomly selected schools in Osun State. The sample size is considered adequate following the proposed of over 500 needed while using IRT model by Embretson and Reise (2000). The instrument for the study consisted of students' responses to 60 multiple-choice Mathematics examination. The responses were dichotomously scores as “1” for correct option and “0” for incorrect option. During the data filtering, omitted options were replaced with “0” and 'Items Not Reach' was replaced with “N”. This was necessary to avoid possible error terms during item calibration for 2-parameter estimates (difficulty **'b',** and discrimination, reliability coefficient, and item characteristic curve (ICC). The 2-parameter model was used, since it has been conformed to provide a better evidence of ICC (Ojerinde; Popoola; Ojo; & Onyeneho, 2012). Thereafter, two separate files (control and data files) were generated into notepad for the analysis using X-Calibre 4.2 package.

Results

What are the item parameter estimates of the Mathematics multiple-choice items using 2-parameter model?

To answer this question, the 60 Mathematics multiple-choice items were calibrated under 2-parameter model of IRT to determine item difficulty (b) and item discrimination (a). The result is presented in Table 1.

Table 1: Item Parameters (difficulty and discrimination) for 60 Calibrated Item under 2-Parameter Model

Items	Item Difficulty (b)	Item Discrimination (a)	Items	Item Difficulty (b)	Item Discrimination (a)
1	0.26	0.07	31	2.86	0.12
2	-1.37	0.48	32	3.24	0.16
3	3.11	0.11	33	-0.94	0.86
4	2.97	0.08	34	2.56	0.14
5	-0.27	0.67	35	3.95	0.23
6	4.00	0.13	36	4.00	0.27
7	-0.21	0.78	37	4.00	0.16
8	-1.07	0.82	38	3.63	0.23
9	0.03	0.57	39	-0.86	0.82
10	2.09	0.08	40	4.00	0.11
11	1.09	0.07	41	-0.15	0.62
12	4.00	0.13	42	-0.61	1.10
13	-1.80	0.70	43	-1.25	1.17
14	-1.15	0.69	44	-1.63	0.54
15	-1.67	0.71	45	-0.19	0.81
16	-2.76	0.41	46	4.00	0.16
17	-0.86	0.91	47	-0.65	0.73
18	-2.17	0.56	48	3.84	0.26
19	-0.77	0.73	49	-0.81	0.84
20	-0.99	0.62	50	4.00	0.12
21	-0.92	0.71	51	4.00	0.28
22	-1.54	0.83	52	-1.43	0.49
23	-1.35	0.57	53	-0.83	1.02
24	3.07	0.13	54	4.00	0.22
25	-1.54	0.52	55	-1.05	1.03
26	0.18	0.52	56	-1.59	0.58
27	-1.29	0.71	57	-0.61	0.76
28	-0.52	0.94	58	-0.15	0.80
29	3.03	0.15	59	-0.96	0.63
30	4.00	0.25	60	-0.83	1.17

Table 1 shows the difficulty (b) and discrimination (a) indices for the 60 Mathematics multiple-choice calibrated items. The indices range from 4.00 to -2.76 and 1.17 to 0.07 for difficulty and discrimination respectively. To identify items that of are faulty and good qualities, a summary Table is presented in Table 2, as proposed by Georgiev, 2008.

Table 2: Distributions of Item Difficulty and Item Discrimination Indices

Item Difficulty (b)	N	Item Discrimination (a)	N
Easy (- 3.00≤b≥ - 1.00)	16(26.7%)	Poor (a≤0.34)	23(38.3%)
		Marginal (0.35.00≤a≤0.64)	13(21.7%)
Moderate (-1.00≤b≥1.00)	22(36.7%)	Moderate (0.65.00≤a≤1.34)	24(40.0%)
		Good (1.35.00≤a≤1.69)	0(0.00%)
Difficult (1.00≤b≥2.00)	22(36.7%)	Excellent (a≥0.34)	0(0.00%)

Table 2 shows that 16(26.7%) items were easy in terms of its difficulty, 22(36.7%) are moderately difficult and 22(36.7%) were difficult. In terms of discrimination indices, 23(38.3%) of the items discriminate poorly, 13(21.7%) discriminate marginally, 24(40.0%) discriminate moderately and non of the items had good and excellently discrimination.

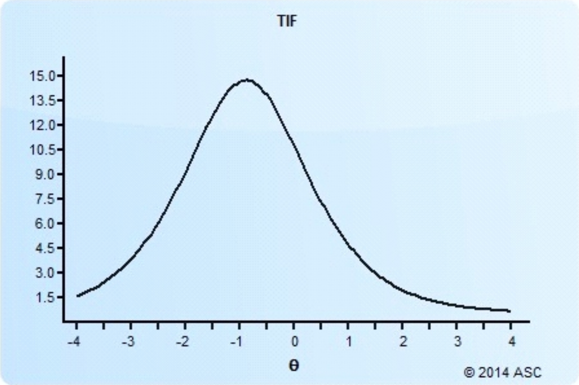


Figure 1 displays a graph of the Test Information Function for all calibrated items. The TIF is a graphical representation of how much information the test is providing at each level of theta. Maximum information was 14.748 at theta = -0.900.

2. What is reliability coefficient of the Mathematics multiple-choice item?

To provide answer to this question, the 60 Mathematics multiple-choice items were calibrated and the results are presented in Tables 2, 3 and 4 respectively.

Table 2: Summary statistics for all calibrated items

Parameter	Items	Mean	SD	Min	Max
a	60	0.5186	0.3249	0.0721	1.1737
b	60	0.6525	2.2407	-2.7568	4

Table 3: Summary statistics for the Total Scores

Test	Items	Alpha	Mean	SD	Skew	Min	Q1	Median	Q3	Max	IQR
Full Test	60	0.8501	33.1196	8.3447	-0.3896	4	28	34	38	53	10

Table 4: Summary statistics for the Theta Estimates

Test	Examinees	Mean	SD	Skew	Min	Q1	Median	Q3	Max	IQR
Full Test	2500	-0.0119	0.9894	-0.0375	-3.4942	-0.6978	0.0071	0.6704	3.7495	1.3682

Table 2 reveals the summary statistics of all calibrated items with **a** ($\overline{X}$=0.5186; SD=0.3249); and **b** ($\overline{X}$=0.6525, SD=0. 2.2407). Also, in Table 3, the classical statistics yielded $\overline{X}$ = 33.12 SD= 8.34, with reliability coefficient of 0.85 of the 60-item. This implied that the Mathematics multiple-choice items had appropriate internal consistency. Finally, Table 4 reveals the summary statistics for the theta estimate with =$\overline{X}$-0.0119and SD=0.9894.

To answer this question, the students' responses to the 60 Mathematics multiple-choice items were calibrated under the 2-parameter model to produce item characteristics curves which show the relationship between students' performance in Mathematics and the characteristic underlying item performance in a monotonically S-Shape. Examples of some of the Mathematics multiple-choice items that comply with the ICC with their respective statistical information are presented in Figures1 to 4

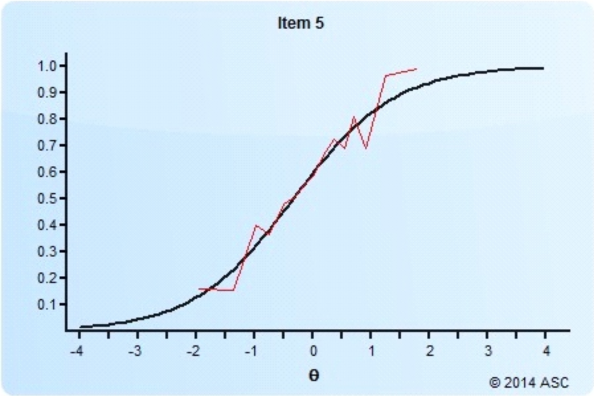
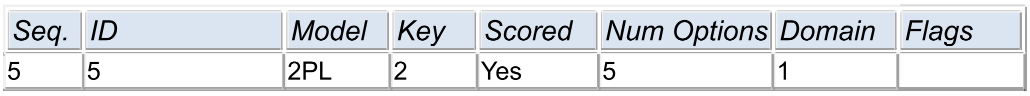


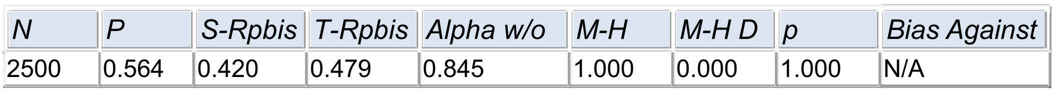
Figure 1: Item Characteristic Curve for Item 5 under 2-parameter models of IRT

Item information



Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
5	5	2PL	2	Yes	5	1	

Classical statistics



N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*	*M-H*	*M-H D*	*p*	*Bias Against*
2500	0.564	0.420	0.479	0.845	1.000	0.000	1.000	N/A

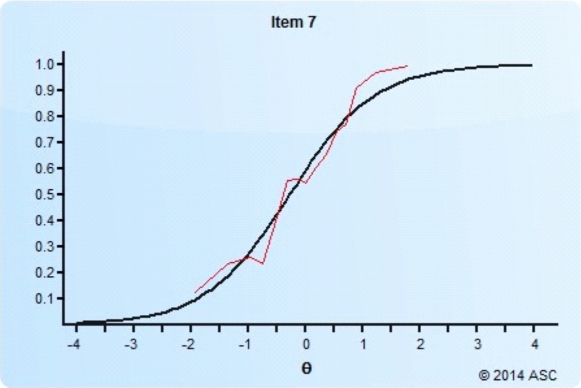
IRT parameters

a	b	a SE	b SE	Chi-sq	df	p	z Resid	p
0.669	-0.272	0.048	0.040	55.600	13	0.000	0.761	0.446

Option statistics

Option	N	Prop.	S-Rpbis	T-Rpbis	Mean	SD	
A	305	0.122	-0.127	-0.235	-0.635	0.761	
B	1410	0.564	0.420	0.479	0.405	0.878	**KEY**
C	202	0.081	-0.177	-0.232	-0.785	0.908	
D	55	0.022	-0.082	-0.102	-0.683	0.775	
E	449	0.180	-0.225	-0.167	-0.366	0.881	
Omit	79	0.032	-0.115	-0.106	-0.593	0.772	
Not Admin	0						

Figure 2: Item Characteristic Curve for Item 5 under 2-parameter models of IRT



Item information

Seq.	ID	Model	Key	Scored	Num Options	Domain	Flags
7	7	2PL	5	Yes	5	1	

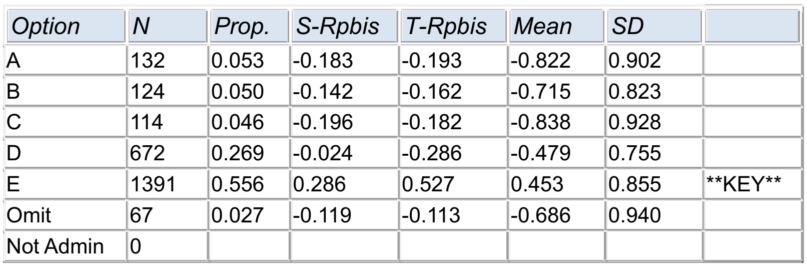
Classical statistics

b	t	{ -Rpbis	Ç-Rpbis	! ▄♫╙Ă ŎǬ	a -H	a -H D	♫	. ╢Ăℓ ! ┼Ă╙▄ℓ Ċ
وﻴﻬﻬ	ﻬ↑ﻴﻴﻱ	ﻬ↑ ﻮﻳ ﻱ	ﻬ↑ﻴﻮﻳ	ﻬ↑ ﻳﻰ	ﺍ↑ﻬﻬﻬ	ﻬ↑ﻬﻬﻬ	ﺍ↑ﻬﻬﻬ	b !

IRT parameters

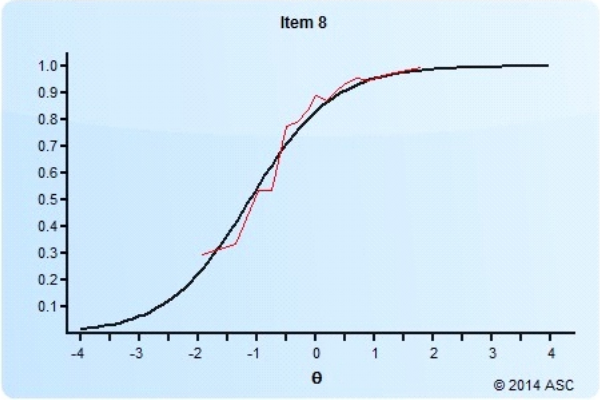
a	b	a SE	b SE	Chi-sq	df	p	z Resid	p
0.779	-0.213	0.045	0.035	50.528	13	0.000	0.957	0.339

Option statistics

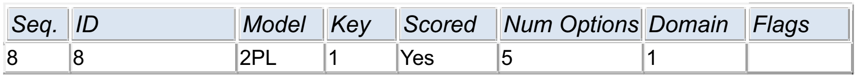


Option	N	Prop.	S-Rpbis	T-Rpbis	Mean	SD	
A	132	0.053	-0.183	-0.193	-0.822	0.902	
B	124	0.050	-0.142	-0.162	-0.715	0.823	
C	114	0.046	-0.196	-0.182	-0.838	0.928	
D	672	0.269	-0.024	-0.286	-0.479	0.755	
E	1391	0.556	0.286	0.527	0.453	0.855	**KEY**
Omit	67	0.027	-0.119	-0.113	-0.686	0.940	
Not Admin	0						

Figure 3: Item Characteristic Curve for Item 5 under 2-parameter models of IRT

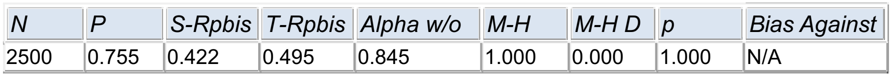


Item information



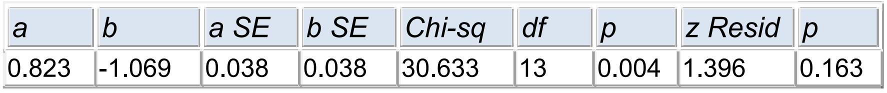
Seq.	ID	Model	Key	Scored	Num Options	Domain	Flags
8	8	2PL	1	Yes	5	1	

Classical statistics



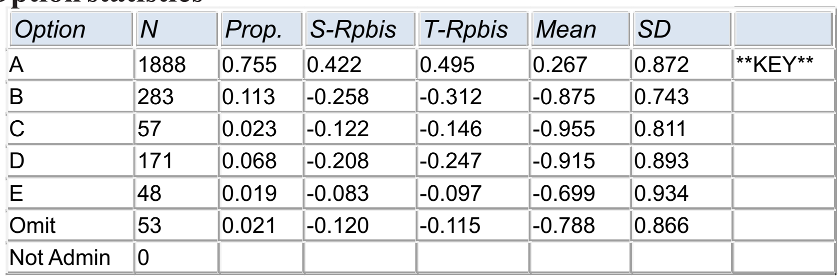
N	P	S-Rpbis	T-Rpbis	Alpha w/o	M-H	M-H D	p	Bias Against
2500	0.755	0.422	0.495	0.845	1.000	0.000	1.000	N/A

IRT parameters



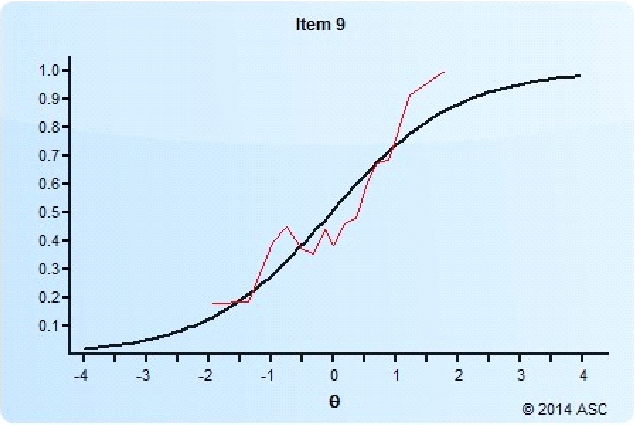
a	b	a SE	b SE	Chi-sq	df	p	z Resid	p
0.823	-1.069	0.038	0.038	30.633	13	0.004	1.396	0.163

Option statistics

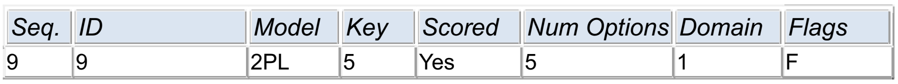


Option	N	Prop.	S-Rpbis	T-Rpbis	Mean	SD	
A	1888	0.755	0.422	0.495	0.267	0.872	**KEY**
B	283	0.113	-0.258	-0.312	-0.875	0.743	
C	57	0.023	-0.122	-0.146	-0.955	0.811	
D	171	0.068	-0.208	-0.247	-0.915	0.893	
E	48	0.019	-0.083	-0.097	-0.699	0.934	
Omit	53	0.021	-0.120	-0.115	-0.788	0.866	
Not Admin	0						

Figure 3: Item Characteristic Curve for Item 5 under 2-parameter models of IRT

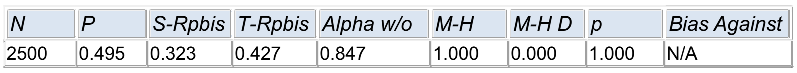


Item information



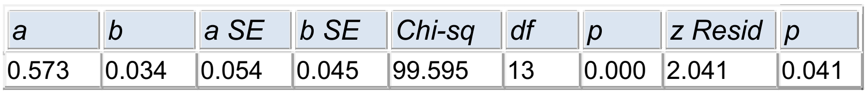
Seq.	ID	Model	Key	Scored	Num Options	Domain	Flags
9	9	2PL	5	Yes	5	1	F

Classical statistics



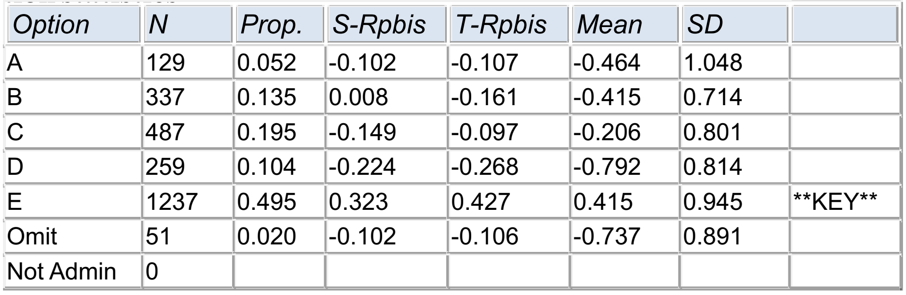
N	P	S-Rpbis	T-Rpbis	Alpha w/o	M-H	M-H D	p	Bias Against
2500	0.495	0.323	0.427	0.847	1.000	0.000	1.000	N/A

IRT parameters



a	b	a SE	b SE	Chi-sq	df	p	z Resid	p
0.573	0.034	0.054	0.045	99.595	13	0.000	2.041	0.041

Option statistics



Option	N	Prop.	S-Rpbis	T-Rpbis	Mean	SD	
A	129	0.052	-0.102	-0.107	-0.464	1.048	
B	337	0.135	0.008	-0.161	-0.415	0.714	
C	487	0.195	-0.149	-0.097	-0.206	0.801	
D	259	0.104	-0.224	-0.268	-0.792	0.814	
E	1237	0.495	0.323	0.427	0.415	0.945	**KEY**
Omit	51	0.020	-0.102	-0.106	-0.737	0.891	
Not Admin	0						

The Figures show ICC for 4 items (5, 7, 8, And 9) and their respective statistical information. The results revealed that thirty-two (32) items representing 53% complied with the assumption of item characteristic curve under 2-parameter model and were adjudged to be good items.

Discussion of Findings

This study focused on the analysis of item parameter estimates of the Mathematics Multiple-choice items used by NECO during 2012 examination period. Specifically, item difficulty (b) and discrimination (a) were considered under 2-parameter model. With reference to the item difficulty, less than 50% of the items fell within the acceptable range of item difficulty, as proposed by Surachi and Rana, (2014). This may likely be attributed to a number of factors such as item ambiguity, item miskeyed, lack of understanding of the questions, improper preparation on the part of the students. Other personal variables may also come to play (Bichi, 2015). In term of discrimination, more than 50% of the items were within the acceptable range of discrimination index suggested by Georgiev, (2008).The result of the research question two revealed that the 60 calibrated Mathematics multiple-choice items is reliable, since, it yielded internal consistency reliability of 0.85. This was higher than the proposed reliability coefficient of 0.70 by Nunnally (1978). The third question revealed that a little above 50% of the items satisfied the assumption of ICC under 2-parameter model.

Conclusion

The study therefore concluded that the Mathematics multiple-choice items were reliable, but further analysis revealed some deficiencies in the items in terms of difficulty, discrimination and compliance with the ICC under 2-parameter model.

Recommendations

Arising from the findings of this study, the following recommendations are made;

(i) More attention should be placed on the use of modern statistical tools for analyzing responses from students after test administration.

(ii) The 3-parameter model can further be use to ascertain or reject the claims under the 2-parameter model.

References

Bichi, A. A. (2015). Item Analysis using a Derived Science Achievement Test Data. *International Journal of Science and Research (IJSR.). Vol. 4(5),* 1656-1662.

Embretson, S. E. and Reise, S. P. (2000). Item response theory for psychologist. Erlbaum, Mahwah, NJ.

Georgiev, N. (2008). Item Analysis of C, D and E Series from Raven's Standard Progressive Matrices with Item Response Theory Two Parameter Logistics Model. *Europe's Journal of Psychology. Vol. 4(3)*

Gocken, M. E. (2014). Math and its Use in Everyday Activity. Retrieved from https://en.wikiversity.org/wiki/Math_and_its_Use_in_Everyday_Activity on 2nd November, 2017. Ikitde, G. A., & Udoh, N.

Hambleton, R. K. & Jones, R. W. (1993). Comparison of Classical Test Theory and Item Response Theory and their Application to Test Development. *Educational Measurement: Issues and Practice, 12(3),* 38-47.

Jimoh, M. I., Daramola, D. S., Oladele, J. I., & Sheu, A. L. (2020). Assessment of Items Prone to Guessing in SSCE Economics Multiple-Choice Tests among Students in Kwara State, Nigeria. Anatolian Journal of Education, 5(1), 17-28.
https://eric.ed.gov/?id=EJ1249146

Klein, S. P. , & Hamilton, L. S. (1999). Large-scale testing Curren practices and new directions (IP-182). Santa Monica, CA: RAND

Nunnally, J. C. (1978). *Educational Measurement and Evaluation. 2nd edition.* New York McGraw-Hill.

Odukoya, J. A., Adekeye, O., Igbinoba, A.O., et al. Item analysis of university wide multiple choice objective examinations: the experience of a Nigerian private University. Qual Quant 52, 983–997 (2018). https://doi.org/10.1007/s11135-017-0499-2

Ojerinde, A., Popoola, O., and Ajeigbe, T. O. (2020). Application of Item Response Theory in Educational Assessment in Africa. Lambert Academic Publishing.

Orlando, M., Sherbourve, C. D. & Thissen, D (2001). Summed-score linking using Item Response Theory: Application to depression measured. Psychological Assessment, 12(3), 354 – 359.

Saeed, R. R. and Noor, M. (2011). Manual on Test Item Construction Technique. Retrieved on

Suruchi, & Rana, S. R. (2014). Test Item Analysis and Relationship between Difficulty Level and Discrimination Index of Test Items in an Achievement Test in Biology. *Paripex-Indian Journal of Research. Vol. 3(6)* 56-58.